Table of Contents

1. Introduction

1.1 Strong Artificial Intelligence

Narrow (or weak) AI is the kind of artificial intelligence that does well at a very narrow range of tasks for which it was designed. Its defining characteristic is its rigidity. New narrow AI algorithms and implementations have to be created and/or trained for each new type of problem or situation we wish to automate. Further, there are many conscious and unconscious processes that humans take for granted that can not be attempted by any narrow AI, neither now nor in the future. This is not due to degrees of effectiveness but a fundamental difference in kind.

Narrow AI represents a fundamental misunderstanding of the role of conscious processing in the derivation of value and meaning. This is not just a philosophical conundrum, but a very practical and scientific matter that impacts its construction, effectiveness, and efficiency. Current approaches, including deep learning and other popular methods, are fundamentally incapable of bridging the gap between mere automation and the machine understanding required to achieve the performance and generality to move beyond their own limitations. This is true regardless of computing power or advances in new types of hardware.

By contrast, strong AI will have the capacity to apply past experience to new problems and challenges. Its defining characteristic is its generality (adaptability). Like us, it will have the ability to adjust and operate in new environments or situations with growing effectiveness over time. Unlike narrow AI, it doesn't have to be reprogrammed or redesigned for each new type of situation or problem it attempts to solve. Most importantly, however, will be its ability to understand meaning and derive value, which presuppose higher cognition in both machines and animals. In addition to these abilities, strong AI will also be vastly more efficient, as it will not have to crudely approximate the benefits of machine understanding through brute-association and enumeration.

Strong AI represents the *ne plus ultra* of human achievement; there is simply nothing more beyond this in terms of impact. Once achieved, we will have unlocked the secrets of abstract cognition, enabling us to do labor and research that will be limited only by the material and energy resources we choose to pool towards it. The eradication of poverty, hunger, and disease will be virtually assured. Humanity will have realized the means to achieve its greatest ambitions and dreams. But not without cost.

1.2 Motivation

The immediate threat will not be from strong AI itself, but from those who will utilize it. Strong AI is a force multiplier. It enhances the power and effectiveness of that which is used in conjunction with it, and there is no realistic and practical way to dictate who uses this power in the world once released. Further, it will not be possible to prevent its eventual release nor limit its spread. Laws and regulation will be ineffective the same way they have been ineffective at combating the piracy of various digital works. The difference being that it will only take one successfully reverse engineered copy of strong AI for the threat model to change permanently. In the end, anyone who wants access to strong AI in the future will eventually gain

access to it regardless of any and all restrictions we build into it or around it.

There is also now an emerging threat from misinformation and propaganda. These campaigns are leading people to believe that there are solutions to problems in the safety, ethics, and security of artificial intelligence that will not, in fact, exist due to technical and practical limitations. They promulgate fear and seek to delay the development and use of this technology in an attempt at leveraging political power, profiteering from sensationalism, or in a vain effort to protect society by attempting to control the inherently uncontrollable medium in which strong AI will be used. These public displays, open letters, and gestures are inert and the initiatives behind them are incapable of addressing the security issues raised in this book. Tamper resistance and moral intelligence will be useful methods for day-to-day use of AI implementations involving small numbers of people, but will not thwart misuses of this technology that impact large populations. To make the situation worse, there exists no practical means to stop such abuses that can not be easily circumvented by those with the expertise. This will not change with time or greater ingenuity.

There are three things that motivate this book. The first is to make it crystal clear that we are ultimately powerless to stop the release and eventual abuse of this technology. The second is to show that the best case scenario requires a fundamental change to society, and possibly to human nature itself. With strong AI, we may have reached a point where individual power has exceeded the means of conventional human power structures. When this happens, we will be judged not by some subverting force of super-intelligence, but by our own genetic and cultural baggage. The malevolent among us will have access to limitless knowledge and expertise over any subject, with the means to cause great harm without central organization or large resources.

This leads to the third and final motivation. The most realistic scenario to mitigate the destructive potential of this technology is to actually develop it and instrument it for defensive purposes as soon as possible, before it is developed unexpectedly somewhere in the world. We must cooperate in this game-theoretic step by making this first cooperative move. While this may appear counter-intuitive, it becomes clearer with an acceptance of the inevitable mentioned above.

Realistically, it is extremely improbable that we will dramatically change, let alone as a species, sufficiently enough to be responsible in our use of strong AI, down to the last person, before it arrives. And, based on the present rate of propaganda and politicizing of the issue, we will have invoked the largest moral hazard of all time by constructing a false sense of security. Given the facts, as will be shown throughout this book, the most logical strategy will be to exploit a first-move advantage by developing this technology now and using the only advantage that large power structures have over asymmetric actors: large resources. By developing large defensive strong AI systems, we may be able to stay ahead of malicious users of the technology. This represents the most realistic hope in what will become a developmental struggle for humanity as it learns to cope with a new found power over thought and experience.

▲ Return to Top